Learning Invariant Representation of Tasks for Robust Surgical State Estimation

Yidan Qin[1,2], Max Allan[1], Yisong Yue[2], Joel W. Burdick[2], Mahdi Azizian[1]

***Abstract*—Surgical state estimators in robot-assisted surgery (RAS) - especially those trained via learning techniques - rely heavily on datasets that capture surgeon actions in laboratory or real-world surgical tasks. Real-world RAS datasets are costly to acquire, are obtained from multiple surgeons who may use different surgical strategies, and are recorded under uncontrolled conditions in highly complex environments. The combination of high diversity and limited data calls for new learning methods that are robust and invariant to operating conditions and surgical techniques. We propose *StiseNet*, a Surgical Task Invariance State Estimation Network with an invariance induction framework that minimizes the effects of variations in surgical technique and operating environments inherent to RAS datasets. StiseNet's adversarial architecture learns to separate nuisance factors from information needed for surgical state estimation. StiseNet is shown to outperform state-of-the-art state estimation methods on three datasets (including a new real-world RAS dataset: HERNIA-20).**

***Index Terms*—Surgical Robotics: Laparoscopy, Deep Learning Methods, AI-Based Methods, Medical Robots and Systems.**

I. Introduction

WHILE the number of Robot-Assisted Surgeries (RAS) continues to increase, at present they are entirely based on teleoperation. Autonomy has the potential to improve surgical efficiency and to improve surgeon and patient comfort in RAS, and is increasingly investigated [1]. Autonomy can be applied to passive functionalities [2], situational awareness [3], and surgical tasks [4], [5]. A key prerequisite for surgical automation is the accurate real-time estimation of the current surgical state. Surgical states are the basic elements of a surgical task, and are defined by the surgeon's actions and observations of environmental changes [6]. Awareness of surgical states would find applications in surgical skill assessment [7], shared control, and workflow optimization [8].

Short duration surgical states, with their inherently frequent state transitions, are challenging to recognize, especially in real-time. Many prior surgical state recognition efforts have employed only one type of operational data. Hidden Markov Models [7], [9], Conditional Random Fields (CRF) [10], Temporal Convolutional Networks (TCN) [11], Long-Short Term Memory (LSTM) [12], and others have been used to recognize surgical actions using robot kinematics data. Methods based on Convolutional Neural Networks (CNN), such as CNN-TCN [11] and 3D-CNN [13], have been applied to endoscopic vision data. RAS datasets consist of synchronized data streams. The incorporation of multiple types of data, including robot kinematics, endoscopic vision, and system events (e.g., camera follow: a binary variable indication of if the endoscope is moving), can improve surgical state estimation accuracy in methods such as Latent Convolutional Skip-Chain CRF [14] and Fusion-KVE [6].

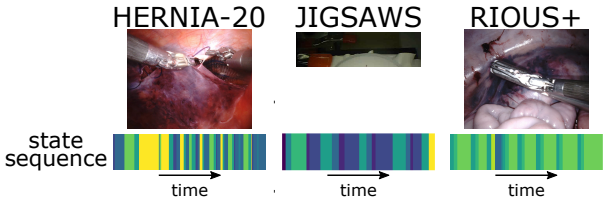

Fig. 1: **Top row:** typical endoscopic images from HERNIA-20 (left), JIGSAWS (middle) and RIOUS+ (right) datasets. **Bottom row:** related surgical state sequence samples, where each color represents a different surgical state.

Prior surgical state estimators relied heavily on RAS datasets for model fitting/training. Limitations in the dataset can be propagated (and perhaps amplified) to the estimator, possibly resulting in a lack of robustness and cross-domain generalizability [14]. Many surgical activity datasets are derived from highly uniform tasks performed using the same technique in only one setting. E.g., the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) [15] suturing task was obtained in a bench-top setting, with suturing performed on marked pads (Fig. 1). Valuable anatomical background visual information is not present in the training data, which may lead to errors when the estimator is applied in real-world surgeries. Moreover, state estimators that are trained on datasets devoid of endoscope motion do not generalize well to new endoscopic views. Endoscope movements are frequent and spontaneous in real-world RAS. Additionally, operators in existing surgical activity datasets typically perform the task with the same technique, or were instructed to follow a predetermined workflow, which limits variability among trials. These limitations can cause state estimators to overfit to the techniques presented during training, and make inaccurate associations between surgical states and specific instrument placements and visual layout, instead of truly relevant features.

In real-world RAS, endoscope lighting and angles, surgical backgrounds, and patient health condition vary considerably among trials, as do state transition probabilities. We consider these variations as potential *nuisance factors* that increase the training difficulty of a robust surgical state estimator. Surgeons

Manuscript received: October 15, 2020; Revised January 11, 2021; Accepted February 16, 2021.

This paper was recommended for publication by Editor Pietro Valdastri upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by Intuitive Surgical Inc.

[1]Intuitive Surgical Inc., 1020 Kifer Road, Sunnyvale, CA,94086, USA
[2]Mechanical and Civil Engineering, Caltech, Pasadena, CA, 91125, USA
Emails: Ida.Qin@intusurg.com, Mahdi.Azizian@intusurg.com

Digital Object Identifier (DOI): see top of this page.

may employ diverse techniques to perform the same surgical task depending on patient condition and surgeon preferences. While the effects of nuisances and technique variations on estimation accuracy can be reduced by a large and diverse real-world dataset, such datasets are costly to acquire.

The combination of limited data and high diversity calls for more robust state estimation training methods, as state-of-the-art methods are not accurate enough for adoption in the safety critical field of RAS. Surgical state estimation can be made invariant to irrelevant nuisances and surgeon techniques if latent representations of the input data contain minimal information about those factors [16]. *Invariant representation learning* (IRL) has been an active research topic in computer vision, where robustness is achieved through *invariance induction* [16]–[20]. Zemel et al. proposed a supervised adversarial model to achieve fair classification under the two competing goals of encoding the input data correctly and obfuscating the group to which the data belongs [17]. A regularized loss function using information bottleneck also induces invariance to nuisance factors [18]. Jaiswal et al. described an adversarial invariance framework in which nuisance factors are distinguished through disentanglement [19], and bias is distinguished through the competition between goal prediction and bias obfuscation [20]. Previous work on IRL via adversarial invariance in time series data focused mostly on speech recognition [21], [22]. RAS data, arising from multiple sources, provides a new domain for IRL of high-dimensional noisy time series data.

Contributions: We propose *StiseNet*, a surgical state estimation model that is largely invariant to nuisances and variations in surgical techniques. StiseNet's adversarial design pits two composite models against each other to yield an invariant latent representation of the endoscopic vision, robot kinematics, and system event data. StiseNet learns a split representation of the input data through the competitive training of state estimation and input data reconstruction, and the disentanglement between essential information and nuisance. The influence of surgeon technique is excluded by adversarial training between state estimation and the obfuscation of a latent variable representing the technique type. StiseNet training does not require any additional annotation apart from surgical states. Our main contributions include:

- Proposing an adversarial model that promotes invariance to nuisance and surgical technique factors in RAS data.
- Designing a process to learn invariant latent representations of real-world RAS data streams, minimizing the effect of factors such as patient condition and surgeon technique.
- Improving frame-wise surgical state estimation accuracy for online and offline real-world RAS tasks by up to 7%, which translates to a 28% relative error reduction.
- Combining semantic segmentation with endoscopic vision to leverage a richer visual feature representation.
- Demonstrating the method on 3 RAS datasets.

StiseNet is evaluated and demonstrated using JIGSAWS suturing [15], RIOUS+ [6], and a newly collected HERNIA-20 dataset containing real-world hernia repair surgeries. StiseNet outperforms state-of-the-art surgical state estimation methods and improves frame-wise state estimation accuracy to 84%. This level of error reduction is crucial for state estimation to gain adoption in RAS. StiseNet also accurately recognizes actions in a real-world RAS task even when a specific technique was not present in the training data.

II. METHODS

StiseNet (Figs. 2 and 3) accepts synchronized data streams of endoscopic vision, robot kinematics, and system events as inputs. To efficiently learn invariant latent representations of noisy data streams, we adopt an adversarial model design loosely following Jaiswal et al. [20] but with model architectures more suitable for time series data. Jaiswal et al.'s adversarial invariance framework for image classification separates useful information and nuisance factors, such as lighting conditions, before performing classification. StiseNet extends this idea by separating learned features from RAS time series data into desired information for state estimation ($\boldsymbol{e}_1$) and other information ($\boldsymbol{e}_2$). Estimation is performed using $\boldsymbol{e}_1$ to eliminate the negative effects of nuisances and variations in surgical techniques. LSTM computational blocks are used for feature extraction and surgical state estimation. LSTMs learn memory cell parameters that govern when to forget/read/write the cell state and memory [12]. They therefore better capture temporal correlations in time series data. StiseNet's components and training procedure are described next. Table I lists key concepts and notation.

A. Feature extraction

Fig. 2 depicts the extraction of features from endoscopic vision, robot kinematics, and system events data. Visual features are extracted by a CNN-LSTM model [24], [25]. To eliminate environmental distractions in the endoscopic view, a previously trained and frozen surgical scene segmentation model based on U-Net [26] extracts a pixel-level semantic mask for each frame. We use two scene classes: tissue and surgical instrument. The semantic mask is concatenated to the unmodified endoscope image as a fourth image channel. This RGB-Mask image $\boldsymbol{I}_t \in \mathbb{R}^{h \times w \times 4}$ is then input to the CNN-LSTM. We implemented a U-Net-style feature map to extract visual features, $\boldsymbol{x}_t^{vis}$, since a condensed surgical scene representation can be taken advantage of by adapting U-Net

Notation	Description
$\boldsymbol{H}$	Concatenated vision, kinematics, and event features $\boldsymbol{H} = \{\boldsymbol{h}^{vis}, \boldsymbol{h}^{kin}, \boldsymbol{h}^{evt}\}$
s	Surgical state
T_{obs}	Observational window size
$\boldsymbol{e}_1$	All factors pertinent to the estimation of s
$\boldsymbol{e}_2$	All other factors (nuisance factors), which are of no interest to goal variable estimation [23]
l	Latent variable (type of surgical technique)
$\overline{d}$	Mean silhouette coefficient quantifying clustering quality
E	Encoder encodes $\boldsymbol{H}$ into $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$
M	Estimator infers s from $\boldsymbol{e}_1$
ψ	Dropout
R	Reconstructer attempts to reconstruct $\boldsymbol{H}$ from $[\psi(\boldsymbol{e}_1), \boldsymbol{e}_2]$
f_1	Disentangler infers $\boldsymbol{e}_2$ from $\boldsymbol{e}_1$
f_2	Disentangler infers $\boldsymbol{e}_1$ from $\boldsymbol{e}_2$
D	Discriminator estimates l from $\boldsymbol{e}_1$

TABLE I: Key variables, concepts, and notation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2021.3063014, IEEE Robotics and Automation Letters

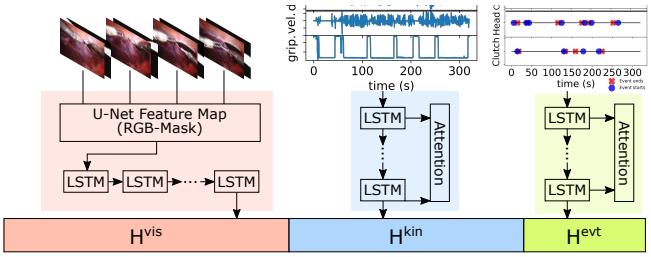

Fig. 2: Features $\boldsymbol{h}^{vis}$, $\boldsymbol{h}^{kin}$, and $\boldsymbol{h}^{evt}$ are respectively extracted from endoscopic vision, robot kinematics, and system events. A semantic mask is appended to the endoscopic vision data to form an RGB-Mask vision input.

weights of the semantic segmentation model trained on a large endoscopic image dataset. We implemented an LSTM encoder to better capture temporal correlations in visual CNN features. This helps the visual processing system to extract visual features that evolve in time. At time t, a visual latent state, $\boldsymbol{h}_t^{vis} \in \mathbb{R}^{n_{vis}}$, is extracted with the LSTM model.

Kinematics data are recorded from the Universal Patient-Side Manipulator (USM) of the da Vinci® Surgical System. Kinematics features are extracted using an LSTM encoder with attention mechanism [27] to identify the important kinematics data types [24]. A multiplier $\boldsymbol{\alpha}_t$, whose elements weight each type of kinematics data, was learned as follows:

$$\boldsymbol{\alpha}_t = softmax\left\{\boldsymbol{u}^T tanh\left(\boldsymbol{W}(\boldsymbol{h}_{t-1}^{kin}, \boldsymbol{c}_{t-1}^{kin}) + \boldsymbol{V}\boldsymbol{X}_t^{kin}\right)\right\}, \quad (1)$$

where $\boldsymbol{h}_{t-1}^{kin}$ is the latent state from the previous frame, $\boldsymbol{c}_{t-1}^{kin}$ is the LSTM cell state, and $\boldsymbol{X}_t^{kin} = (\boldsymbol{x}_{t-T_{obs}+1}^{kin}, ..., \boldsymbol{x}_t^{kin})$ denote the kinematic data inputs. $\boldsymbol{u}$, $\boldsymbol{W}$, and $\boldsymbol{V}$ are learnable parameters. The weighted kinematics data feature vector $\boldsymbol{h}_t^{kin} \in \mathbb{R}^{n_{kin}}$ is calculated as:

$$\boldsymbol{h}_t^{kin} = LSTM(\boldsymbol{h}_{t-1}^{kin}, \boldsymbol{\alpha}_t \cdot \boldsymbol{x}_t^{kin}) \quad (2)$$

The da Vinci® Xi Surgical System also provides system event data (details in Section III). The event features $\boldsymbol{h}_t^{evt}$ are extracted via the same method as kinematics.

B. Feature encoder and Surgical state estimator

As shown in Fig. 3, Encoder E extracts useful information for estimation from the latent feature data $\boldsymbol{H}$. If we assume that $\boldsymbol{H}$ is composed of a set of factors of variation, then $\boldsymbol{H}$ it is composed of mutually exclusive subsets:

- $\boldsymbol{e}_1$: all the factors pertinent to the estimation of the goal variable (the current surgical state s);
- $\boldsymbol{e}_2$: all other factors (nuisance factors), which are of no interest to goal variable estimation [23].

Encoder E is a function trained to partition $\boldsymbol{H}$: $[\boldsymbol{e}_1, \boldsymbol{e}_2] = E(\boldsymbol{H})$. A fully-connected (FC) layer maps $\boldsymbol{H}$ to $\boldsymbol{e}_1$, and another FC layer maps $\boldsymbol{H}$ to $\boldsymbol{e}_2$. Once distinguished, the surgical state s at time t is estimated from the history of the useful signal $\{\boldsymbol{e}_{1,t-T_{obs}+1}, \ldots, \boldsymbol{e}_{1,t}\}$ using an LSTM decoder $\boldsymbol{M}$ following [24]. By learning the parameters in M using $\boldsymbol{e}_1$ instead of $\boldsymbol{H}$, we avoid learning inaccurate associations between nuisance factors and the goal variable.

C. Learning an invariant representation

The invariance induction to nuisance and technique factors is learned via *competition* and *adverseness* between model components [28] (yellow and pink shaded components in Fig. 3). While M encourages the pooling of factors relevant to surgical state estimation in signal $\boldsymbol{e}_1$, a reconstructor R (a function implemented as an FC layer) attempts to reconstruct from the separated signals. Dropout ψ is added to $\boldsymbol{e}_1$ to make it an unreliable source to reconstruct $\boldsymbol{H}$ [19]. This configuration of signals prevents a convergence to the trivial solution where $\boldsymbol{e}_1$ monopolizes all information, while $\boldsymbol{e}_2$ contains none. The mutual exclusivity between $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ is achieved through adversarial training. Two FC layers f_1 and f_2 are implemented as *disentanglers*. f_1 attempts to infer $\boldsymbol{e}_2$ from $\boldsymbol{e}_1$, while f_2 infers $\boldsymbol{e}_1$ from $\boldsymbol{e}_2$. To achieve mutual exclusivity, we should not be able to infer $\boldsymbol{e}_1$ from $\boldsymbol{e}_2$ or vise versa. Hence, the losses of f_1 and f_2 must be maximized. This leads to an adversarial training objective [29]. The loss function with invariance to nuisance factors is:

$$\begin{aligned} L_{nuis} = \alpha L_M\left(s, M(\boldsymbol{e}_1)\right) + \beta L_R\left(\boldsymbol{H}, R(\boldsymbol{e}_2, \psi(\boldsymbol{e}_1))\right) \\ + \gamma\left(L_f\left(\boldsymbol{e}_1, f_1(\boldsymbol{e}_2)\right) + L_f\left(\boldsymbol{e}_2, f_2(\boldsymbol{e}_1)\right)\right) \end{aligned} \quad (3)$$

where α, β, and γ respectively weight the adversarial loss terms [29] associated with architectural components M, R, and disentanglers f_1 and f_2. The training objective with invariance to nuisance factors is a minimax game [28], [30]:

$$\min_{P1} \max_{P2_{nuis}} L_{nuis} \quad (4)$$

where the loss of component $P1 = \{E, M, R\}$ is minimized while the loss of $P2_{nuis} = \{f_1, f_2\}$ maximized.

Besides the presence of nuisance factors, variability in $\boldsymbol{H}$ could also arise from variability in surgical techniques. Variations in technique may not be entirely separable by an invariance to nuisance factors, as they may be correlated to the surgical state. StiseNet therefore adopts an *adversarial debiasing* design [31] that deploys a *discriminator* $D : \boldsymbol{e}_1 \rightarrow l$ for surgical technique invariance. l represents the type of technique employed to perform a surgical task. l is a trial-level categorical attribute that is inferred by k-means clustering of kinematics time series training data based on a dynamic time warping distance metric (function ϕ) [32]. The clusters represent different surgical techniques used in the training trials. The optimal number of clusters k is dataset-specific. To determine it, we implemented the *elbow method* using inertia [33] and the *silhouette method* [34]. The inertia is defined as the sum of squared distances between each cluster member and its cluster center [33] for all clusters. The inertia decreases as k increases, and the elbow point is a relatively optimal k value [33]. The silhouette coefficient d_i for time series i is:

$$d_i = \frac{\underset{j}{mean}\left(\min_{m \notin C_i} \sum_{j \in C_m} \phi(i,j) - \sum_{j \in C_i, j \neq i} \phi(i,j)\right)}{max(a_i, b_i)} \quad (5)$$

where C_i is the cluster of time series i. The operation $\min_{m \notin C_i}$ represents the closest time series to i that does not belong to C_i. We used the mean silhouette coefficient among all time

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2021.3063014, IEEE Robotics and Automation Letters

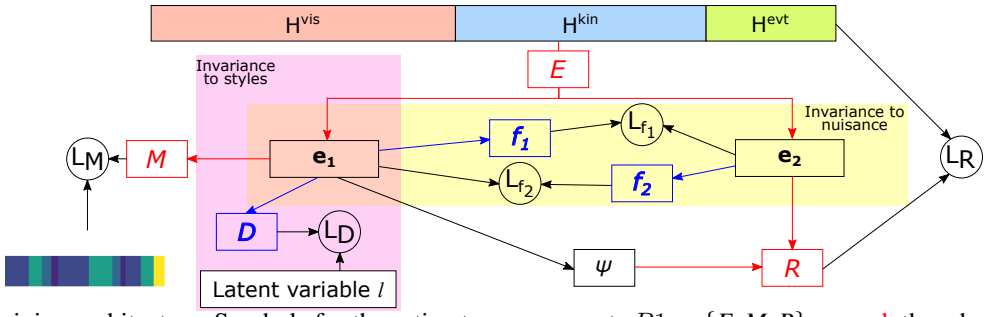

Fig. 3: StiseNet training architecture. Symbols for the estimator components $P1 = \{E, M, R\}$ are red, the adversarial component $P2 = \{f_1, f_2, D\}$ is blue, and training loss calculations are black. $P2$ implements invariance to nuisance (yellow shading) and surgical techniques (pink shading). RAS data features $\boldsymbol{H}$ are divided into information essential for state estimation, $\boldsymbol{e}_1$, and other information $\boldsymbol{e}_2$. $\boldsymbol{H}$ is reconstructed from $\psi(\boldsymbol{e}_1)$ and $\boldsymbol{e}_2$, where ψ is dropout.

series $\overline{d}$ to select k. $\overline{d}$ is a measure of how close each data point in one cluster is to data points in the nearest neighboring clusters. The k with the highest $\overline{d}$ is the optimal number of clusters. The loss function with invariance to both nuisance and surgical techniques is then:

$$L = L_{nuis} + \delta L_D\left(l, D(\boldsymbol{e}_1)\right) \tag{6}$$

where δ is the weight associated with the discriminator loss. The term $P2$ contains an additional term: $\tilde{P2} = \{f_1, f_2, D\}$:

$$\min_{P1} \max_{\tilde{P2}} L. \tag{7}$$

D. Training and inference

StiseNet's feature extraction components were trained following [6]. Specifically, the first three channels of the top layer in U-Net visual feature map were initialized with the weights from the surgical scene segmentation model. The visual input was resized to $h = 256$ and $w = 256$. The extracted features have dimensions $n_{vis} = 40$, $n_{kin} = 40$, and $n_{evt} = 4$, which were determined using grid search. All data sources are synchronized at 10Hz with $T_{obs} = 20\ samples = 2sec.$ The optimal cluster number, k, for JIGSAWS, RIOUS+, and HERNIA-20 were 9, 7, and 4, respectively. The temporal clustering process was repeated to ensure reproducibility due to the randomness in initialization. Section IV described how k is determined in these datasets.

StiseNet is trained end-to-end with the minimax objectives (Eq.s 4 and 7). We used the categorical cross-entropy loss for L_M and L_D. L_f and L_R are mean squared error loss. ψ is a dropout [35] with the rate of 0.4, 0.1, and 0.4 for JIGSAWS, RIOUS+, and HERNIA-20, respectively. To effectively train the adversarial model, we applied a scheduled adversarial optimizer [28], in which a training batch is passed to either $P1$ or $P2$ while the other component's weights are frozen. The alternating schedule was found by grid search to be 1:5.

III. Experimental Evaluation

We evaluated StiseNet's performance on the JIGSAWS suturing [15], RIOUS+ [6], and a newly collected HERNIA-20 dataset, respectively. These datasets were annotated with manually determined lists of fine-grained states (Table II).

A. Datasets

JIGSAWS: The JIGSAWS [15] bench-top suturing task includes 39 trials by eight surgeons partaking in nine surgical actions. We used the endoscopic vision and USM's kinematics (gripper angle, translational and rotational positions and velocities) data. There was no system events data. The tooltips' orientation matrices were converted to Euler angles.

RIOUS+: The RIOUS+ dataset, introduced in [6], [24], captures 40 trials of an ultrasound scanning task on a da Vinci Xi® Surgical System by five users in a mixture of bench-top (27) and OR (13) trials. Eight states represent user actions or environmental changes. Endoscopic vision, USM kinematics, and six binary system events serve as inputs–see [6]. A finite state machine model of the task was determined prior to data collection. The operators were instructed to strictly follow this predetermined task workflow and to ignore environmental disruptions. The action sequences and techniques are therefore highly structured and similar across trials. While it includes more realistic RAS elements, such as OR settings and endoscope movements, RIOUS+ lacks the behavioral variability of real-world RAS data.

HERNIA-20: The HERNIA-20 dataset contains 10 fully anonymized real-world robotic transabdominal preperitoneal inguinal hernia repair procedures performed by surgeons on da Vinci Xi ® Surgical Systems. For performance evaluation, we selected a running suturing task performed to re-approximate the peritoneum, which contains 11 states. The endoscopic vision, USM kinematics, and system events are used as inputs. Because HERNIA-20 captures real-world RAS performed on patients, the robustness of surgical state estimation models can be fully examined.

B. Metrics

The quality of the learned invariant representations of surgical states $\boldsymbol{e}_1$ and other information $\boldsymbol{e}_2$ is visually examined. Arrays of $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ in each state instance (a consecutive block of time frames of the same surgical state) are embedded in 2D space using the Uniform Manifold Approximation and Projection (UMAP) algorithm [36] - a widely-adopted dimension reduction and visualization method that preserves more of the global structure of the data.

We used the percentage of accurately identified frames in a test set to evaluate each model's surgical state estimation

Gesture	**JIGSAWS Suturing Dataset**	Duration (s)
G1	Reaching for the needle with right hand	2.2
G2	Positioning the tip of the needle	3.4
G3	Pushing needle through the tissue	9.0
G4	Transferring needle from left to right	4.5
G5	Moving to center with needle in grip	3.0
G6	Pulling suture with left hand	4.8
G7	Orienting needle	7.7
G8	Using right hand to help tighten suture	3.1
G9	Dropping suture and moving to end points	7.3
State	**RIOUS+ Dataset**	Duration (s)
S1	Probe released, out of endoscopic view	6.3
S2	Probe released, in endoscopic view	7.6
S3	Reaching for probe	3.1
S4	Grasping probe	1.1
S5	Lifting probe up	2.4
S6	Carrying probe to tissue surface	2.3
S7	Sweeping	5.1
S8	Releasing probe	1.7
State	**HERNIA-20 Dataset**	Duration (s)
S1	Reaching for the needle	3.9
S2	Positioning the tip of the needle	3.3
S3	Pushing needle through the tissue	4.2
S4	Pulling tissue with left hand	3.6
S5	Transferring needle from left to right	3.7
S6	Orienting needle	6.6
S7	Pulling suture with left hand	5.8
S8	Pulling suture with right hand	4.8
S9	Transferring needle from right to left	4.6
S10	Using right hand to tighten suture	4.3
S11	Adjusting endoscope	3.8

TABLE II: Datasets State Descriptions and Mean Duration

accuracy. Model performance was evaluated in non-causal and causal settings. In a non-causal setting, the model can use information from future time frames, which is suitable for post-operative analyses. In causal settings, the model only has access to the current and preceding time frames. Surgical state estimation is harder in the causal setting; however, it is a more useful evaluation metric for real-time applications.

We used the source code provided by the authors of the comparison methods when the model performance of a particular setting or dataset was not available [11], [12] and performed training and evaluation ourselves. JIGSAWS suturing and RIOUS+ datasets were evaluated using *Leave One User Out* (LOUO) [15], while HERNIA-20 was evaluated using 5-fold cross validation, since each trial's surgeon ID is not available due to privacy protection.

C. Ablation Study

We compared StiseNet against its ablated versions: StiseNet-Non Adversarial (StiseNet-NA), StiseNet-Nuisance Only (StiseNet-NO), and StiseNet-Technique Only (StiseNet-TO). StiseNet-NA omits the adversarial component P2 entirely and uses $\boldsymbol{H}$ directly for estimation. StiseNet-NO separates useful information and nuisance factors, but excludes the invariance to surgical techniques (pink-shaded area in Fig. 3). StiseNet-TO includes the invariance to techniques but omits the separation between $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$. The ablation study demonstrates the necessity of the adversarial model design and individual contributions of each model component towards a more accurate surgical state estimation.

IV. Results and Discussions

Fig. 4 plots for each dataset the *total inertia* and the *mean silhouette coefficient* $\overline{d}$ as functions of the number of clusters k. Fig. 5 shows the UMAP visualizations of $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ for all surgical states. We compare both the non-causal (Table III) and causal (Table IV) performance of StiseNet with its ablated versions and prior methods. Fig. 6 shows the variability in HERNIA-20 data through sample sequences from three technique clusters, each performed in a distinctively different style with environmental variances. Invariance of StiseNet to nuisances and surgical techniques is shown by its accurate surgical state estimations in the presence of visibly diverse input data. Fig. 7 shows a sample state sequence from HERNIA-20 and the causal state estimation results using multiple methods, including forward LSTM [12], Fusion-KVE [6], and the ablated and full versions of StiseNet.

As mentioned in Section II-C, the optimal number of clusters k can be estimated from the elbow point of the inertia-k curve, or the k associated with the maximum mean silhouette coefficient $\overline{d}$. We implemented both methods and illustrate our choices of k in Fig. 4. The optimal k is easily identifiable for JIGSAWS and HERNIA-20 (Fig. 4a and 4c), with the largest $\overline{d}$ occurs near the "elbow" of the inertia-k curve. A peak in the RIOUS+ mean silhouette coefficient curve is less evident (Fig. 4b). The optimal number of clusters need not match the number of operators, as the inter-personal characteristics are not the only accountable factor for the variations among trials. Intra-personal variations can affect clustering. E.g., JIGSAWS contains metadata corresponding to expert ratings of each trial [15]: the ratings fluctuate among trials performed by the same surgeon. The optimal k determined by kinematics data is somewhat robust against patient anatomy; however, a highly unique patient anatomy can lead surgeons to modify their maneuvers significantly. Such a trial could fall into a different technique cluster.

Fig. 5 visualizes the 2D projections of $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$. The first row shows that $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ separate neatly into two clusters for

Non-causal				
	Input data	JIGSAWS	RIOUS+	HERNIA-20
TCN [11]	kin	79.6	82.0	72.1
TCN [11]	vis	81.4	62.7	61.5
Bidir. LSTM [12]	kin	83.3	80.3	73.8
LC-SC-CRF [14]	vis+kin	83.5	-	-
3D-CNN [13]	vis	84.3	-	-
Fusion-KVE [6]	vis+kin+evt	86.3	**93.8**	78.0
StiseNet-NA	vis+kin+evt	86.5	93.1	80.0
StiseNet-TO	vis+kin+evt	88.1	88.9	81.8
StiseNet-NO	vis+kin+evt	87.9	90.3	83.2
StiseNet	vis+kin+evt	**90.2**	92.5	**84.1**

Table III: State estimation performance in non-causal setting. JIGSAWS results did not include system events.

Causal				
	Input data	JIGSAWS	RIOUS+	HERNIA-20
TCN [11]	vis	76.8	54.8	58.3
TCN [11]	kin	72.4	78.4	68.1
Forward LSTM [12]	kin	80.5	72.2	69.8
3D-CNN [13]	vis	81.8	-	-
Fusion-KVE [6]	vis+kin+evt	82.7	89.4	75.7
StiseNet-NA	vis+kin+evt	83.4	88.9	77.3
StiseNet-TO	vis+kin+evt	84.2	87.1	81.4
StiseNet-NO	vis+kin+evt	84.1	88.9	81.0
StiseNet	vis+kin+evt	**85.6**	**89.5**	**82.7**

Table IV: State estimation performance in a causal setting. JIGSAWS results did not include system events.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2021.3063014, IEEE Robotics and Automation Letters

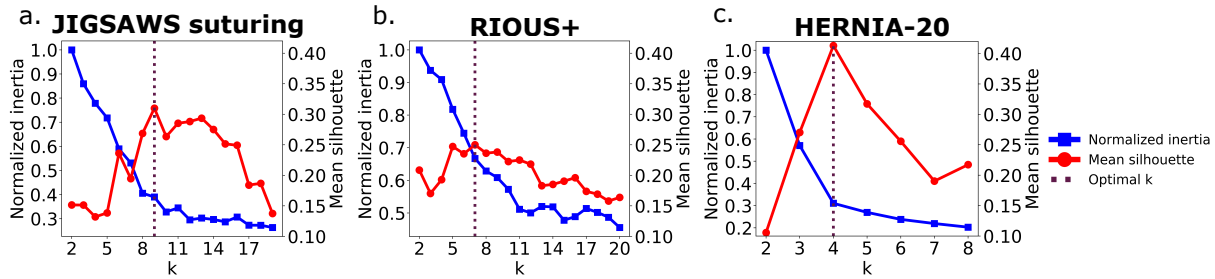

Fig. 4: Normalized inertia (with respect to the maximum value) and mean silhouette coefficient as functions of the number of clusters k for each dataset. The vertical dotted line indicates the optimal k (the maximum mean silhouette coefficient).

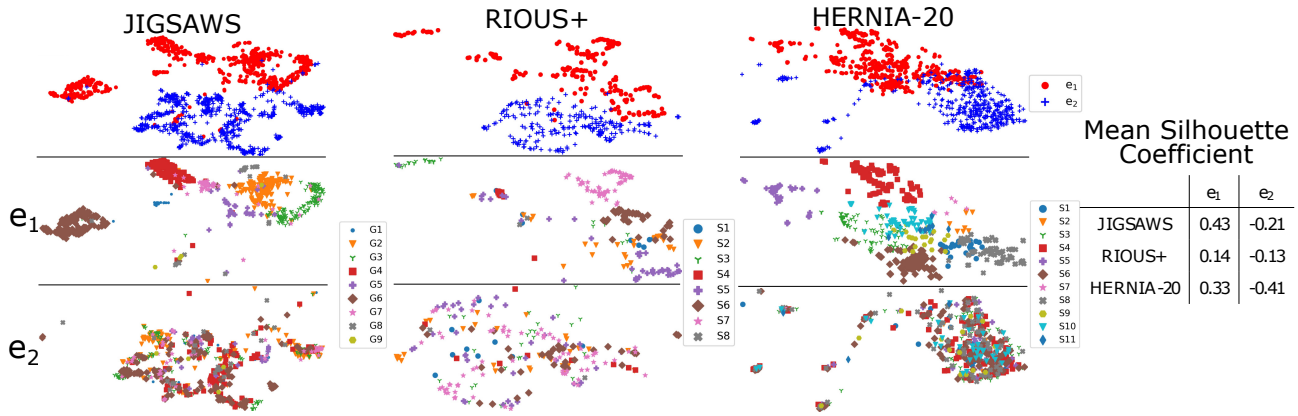

Fig. 5: 2D UMAP plots of information enclosed in $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ at each state instance. **Top row:** $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ segregates into distinguishable clusters, which indicates little overlap in information. **Middle row:** Information in $\boldsymbol{e}_1$ color-coded by surgical states clusters relatively neatly. **Bottom row:** Information in $\boldsymbol{e}_2$ is more intertwined and non-distinguishable by state. The mean silhouette coefficient $\overline{d}$ of each graph is shown, with a larger $\overline{d}$ indicating better clustering quality.

all datasets, validating the effectiveness of disentanglers f_1 and f_2 since $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ contain little overlapping information. Since $\boldsymbol{e}_1$ contains useful information for state estimation, while $\boldsymbol{e}_2$ does not, $\boldsymbol{e}_1$ should be better segregated into clusters associated to each state. The second and third rows of Fig. 5 (color-coded by surgical state) show cleanly segregated clusters for $\boldsymbol{e}_1$, while the $\boldsymbol{e}_2$ projections are not distinguishable by state. The *mean silhouette coefficient* for each graph also supports this observation. This strongly suggests that each surgical state has a unique representation in $\boldsymbol{e}_1$, while $\boldsymbol{e}_2$ contains little information useful for state estimation.

Table III and IV show non-causal and causal surgical state estimation performance of recently proposed methods and StiseNet (and its ablated versions). StiseNet yields an improvement in frame-wise surgical state estimation accuracy for JIGSAWS suturing (up to 3.9%) and HERNIA-20 (up to 7%) under both settings, which shows the necessity and effectiveness of the adversarial model design. The non-causal performance of StiseNet on RIOUS+ is slightly worse compared to our Fusion-KVE method [6], which does not dissociate nuisance or style variables. This result can be explained by StiseNet's model design and training scheme. The added robustness of StiseNet against variations in background, surgical techniques, etc. comes at the cost of the increased training complexity associated with adversarial loss functions and minimax training. Surgeon techniques and styles vary in JIGSAWS, and more significantly in HERNIA-20. Nuisance factors (tissue deformations, endoscopic lighting conditions and viewing angles, etc.) also vary considerably among trials and users in HERNIA-20. However, since RIOUS+ users were instructed to strictly follow a predetermined workflow, there are few nuisance and technique factors. The disentanglement between essential information $\boldsymbol{e}_1$ and other information $\boldsymbol{e}_2$ is therefore less effective. This hypothesis is supported by the observation that the dropout rate required for StiseNet training covergence is 0.1 for RIOUS+, whereas JIGSAWS and HERNIA-20 training converged with a dropout rate of 0.4. A lower dropout rate indicates that $\boldsymbol{e}_2$ contains little information despite the dropout's effort to avoid the trivial solution. Additionally, the uniformity across RIOUS+ participants results in a nearly constant mean silhouette coefficient (Fig. 4b). StiseNet's invariance properties cannot be fully harnessed, explaining its less competitive performance in RIOUS+ as compared to the real-world data of HERNIA-20.

In real-world RAS, surgeons may use different techniques to accomplish the same task. Fig. 6 shows three HERNIA-20 trials with distinctive suturing geometries: suturing from left to right, from right to left, and back and forth along a vertical seam. These trials fall into three clusters. We show images from instances of states S3, S4, S5, S7 and S8 in each trial. These images of different instances of the same state vary greatly not just in technique and instrument layout, but also in nuisance factors such as brightness and endoscope angles. Yet, StiseNet accurately estimates the surgical states due to its invariant latent representation of the input data.

Fig. 7 demonstrates StiseNet's robustness during rapid and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2021.3063014, IEEE Robotics and Automation Letters

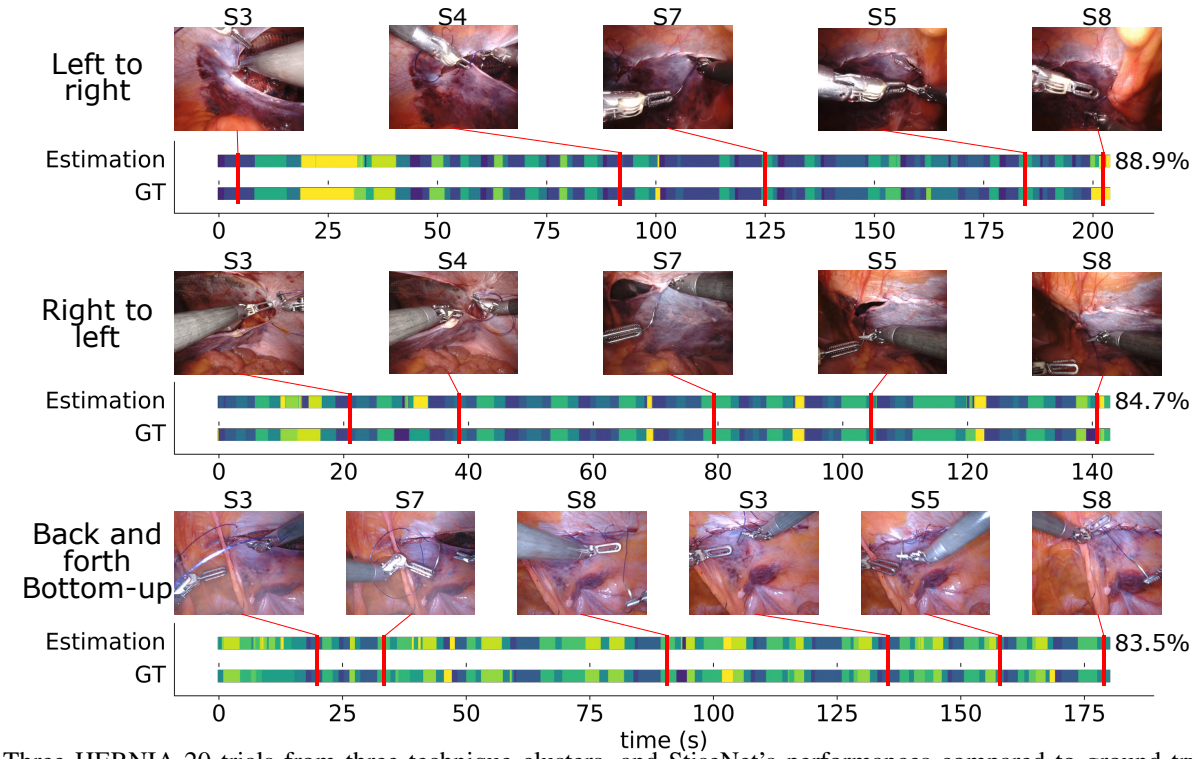

Fig. 6: Three HERNIA-20 trials from three technique clusters, and StiseNet's performances compared to ground truth (GT). Instances of the same state in different trials are substantially and visibly different; however, StiseNet correctly estimates them. Variations across trials arise from both nuisances and techniques. Potential sources of nuisances include but are not limited to lighting conditions, presence of fat or blood, peritoneum color, endoscope movements, etc.

unpredictable state transitions in a real-world RAS suturing task. We compare the causal estimation performance of Forward-LSTM, Fusion-KVE, the ablated, and full versions of StiseNet against ground truth. Forward-LSTM, which only uses kinematics data, has a block of errors from 20s to 30s since it cannot recognize the "adjusting endoscope" state due to a lack of visual and event inputs. When those inputs are added, Fusion-KVE and StiseNet recognize this state. Fusion-KVE still shows a greater error rate due to limited training data with high environmental diversity, which reflects Fusion-KVE's vulnerability to nuisance and various surgical techniques. StiseNet-NO shows fewer error blocks: yet it is still affected by different technique types. The higher estimation accuracy of StiseNet shows its technique-agnostic robustness in real-world RAS, even with a small training dataset that contains behavioral and environmental diversity.

V. Conclusions and Future Work

This paper focused on improving the accuracy of surgical state estimation in real-world RAS tasks learned from limited amounts of data with high behavioral and environmental diversity. We proposed *StiseNet*: an adversarial learning model with an invariant latent representation of RAS data. StiseNet was evaluated on three datasets, including a real-world RAS dataset that includes different surgical techniques carried out in highly diverse environments. StiseNet improves the state-of-the-art performance by up to 7%. The improvement is significant for the real-world running suture tasks, which benefit greatly from invariance to surgical techniques, environments, and patient anatomy. Ablation studies showed the effectiveness of the adversarial model design and the necessity of invariance inductions to both nuisance and technique factors. StiseNet training does not require additional annotation apart from the surgical states. We plan to further investigate alternative labelling methods of surgical techniques and the invariance induction to other latent variables such as surgeon ID, surgeon levels of expertise, etc. Due to the limited data availability, StiseNet has only been evaluated on small datasets. Adding more trials to HERNIA-20 will allow us to evaluate StiseNet more comprehensively. To further improve estimation accuracy, StiseNet's neural network architectures may be further optimized for a better learning of temporal correlations within data. We also plan to incorporate longer-term context information [37], [38]. StiseNet's accurate and robust surgical state estimation could also aide the development of surgeon-assisting functionalities and shared control systems in RAS.

Acknowledgment

We would like to thank Dr. Seyedshams Feyzabadi, Dr. Azad Shademan, Dr. Sandra Park, Dr. Humphrey Chow, and Dr. Wenqing Sun for their support.

References

[1] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, *et al.*, "Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy," *Science Robotics*, vol. 2, no. 4, p. 8638, 2017.

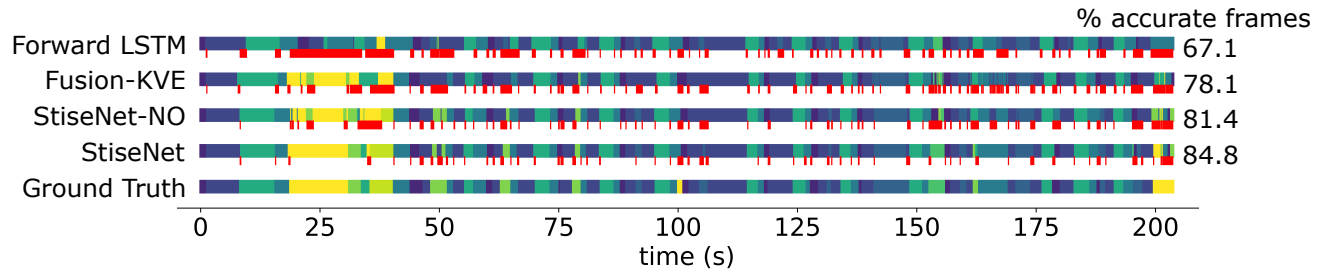

Fig. 7: Example HERNIA-20 surgical state estimation results by forward LSTM [12], Fusion-KVE [6], StiseNet-NO, and StiseNet, compared to ground truth. State estimation results (top) and discrepancies with ground truth in red (bottom) are shown in each block bar.

[2] M. Selvaggio, G. A. Fontanelli, F. Ficuciello, L. Villani, and B. Siciliano, "Passive virtual fixtures adaptation in minimally invasive robotic surgery," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3129–3136, 2018.

[3] P. Chalasani, A. Deguet, P. Kazanzides, and R. H. Taylor, "A computational framework for complementary situational awareness (csa) in surgical assistant robots," in *IEEE Int. Conf. Robotic Computing*, 2018, pp. 9–16.

[4] A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. Kim, "Supervised autonomous robotic soft tissue surgery," *Sci. Rrans. Med.*, vol. 8, no. 337, pp. 337ra64–337ra64, 2016.

[5] A. Attanasio, B. Scaglioni, M. Leonetti, A. F. Frangi, W. Cross, C. S. Biyani, and P. Valdastri, "Autonomous tissue retraction in robotic assisted minimally invasive surgery–a feasibility study," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6528–6535, 2020.

[6] Y. Qin, S. A. Pedram, S. Feyzabadi, M. Allan, A. J. McLeod, J. W. Burdick, and M. Azizian, "Temporal segmentation of surgical sub-tasks through deep learning with multiple data sources," in *IEEE Int. Conf. Robotics and Automation*, 2020, to appear.

[7] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal, "Sparse hidden markov models for surgical gesture classification and skill evaluation," in *Int. conf. info. process. in computer-assisted interven.*, 2012, pp. 167–177.

[8] N. Padoy, "Machine and deep learning for workflow recognition during surgery," *Min. Inv. Ther. & Allied Tech.*, vol. 28, no. 2, pp. 82–90, 2019.

[9] M. Volkov, D. A. Hashimoto, G. Rosman, O. R. Meireles, and D. Rus, "Machine learning and coresets for automated real-time video segmentation of laparoscopic and robot-assisted surgery," in *IEEE Int. Conf. Robotics and Automation*, 2017, pp. 754–759.

[10] L. Tao, L. Zappella, G. D. Hager, and R. Vidal, "Surgical gesture segmentation and recognition," in *Int. Conf. Med. Image Comput. and Computer-Assisted Interven.*, 2013, pp. 339–346.

[11] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conf. Computer Vision*. Springer, 2016, pp. 47–54.

[12] R. DiPietro, C. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee, M. R. Lee, and G. D. Hager, "Recognizing surgical activities with recurrent neural networks," in *Int. Conf. med. image comput. and computer-assisted inter.*, 2016, pp. 551–558.

[13] I. Funke, S. Bodenstedt, F. Oehme, F. von Bechtolsheim, J. Weitz, and S. Speidel, "Using 3d convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video," in *Int. Conf. Med. Image Comp. and Comp.-Assisted Inter.*, 2019, pp. 467–475.

[14] C. Lea, R. Vidal, and G. D. Hager, "Learning convolutional action primitives for fine-grained action recognition," in *IEEE Int. Conf. Robotics and Automation*. IEEE, 2016, pp. 1642–1649.

[15] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, *et al.*, "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *MICCAI Workshop: M2CAI*, vol. 3, 2014, p. 3.

[16] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig, "Controllable invariance through adversarial feature learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 585–596.

[17] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Int. Conf. Machine Learning*, 2013, pp. 325–333.

[18] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," *J. Machine Learning Research*, vol. 19, no. 1, pp. 1947–1980, 2018.

[19] A. Jaiswal, R. Y. Wu, W. Abd-Almageed, and P. Natarajan, "Unsupervised adversarial invariance," in *Advances in Neural Information Processing Systems*, 2018, pp. 5092–5102.

[20] A. Jaiswal, Y. Wu, W. AbdAlmageed, and P. Natarajan, "Unified adversarial invariance," *arXiv preprint arXiv:1905.03629*, 2019.

[21] I. Hsu, A. Jaiswal, P. Natarajan, *et al.*, "Niesr: Nuisance invariant end-to-end speech recognition," *arXiv preprint arXiv:1907.03233*, 2019.

[22] R. Peri, M. Pal, A. Jati, K. Somandepalli, and S. Narayanan, "Robust speaker recognition using unsupervised adversarial invariance," in *IEEE Int. Conf. Acoust., Speech Sig. Proc.*, 2020, pp. 6614–6618.

[23] D. Basu, "On the elimination of nuisance parameters," in *Selected Works of Debabrata Basu*. Springer, 2011, pp. 279–290.

[24] Y. Qin, S. Feyzabadi, M. Allan, J. W. Burdick, and M. Azizian, "davincinet: Joint prediction of motion and surgical state in robot-assisted surgery," *arXiv preprint arXiv:2009.11937*, 2020.

[25] T. Yu, D. Mutter, J. Marescaux, and N. Padoy, "Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition," *arXiv preprint arXiv:1812.00033*, 2018.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[27] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," *arXiv preprint arXiv:1704.02971*, 2017.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Adv. Neural Info. Process. Systems*, 2014, pp. 2672–2680.

[29] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representation using adversarial training," in *Advv Neur, Inf. Process. Syst.s*, 2016, pp. 5040–5048.

[30] M. Maschler, S. Zamir, E. Solan, M. Borns, and Z. Hellman, *Game Theory*. Cambridge University Press, 2013.

[31] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[32] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[33] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*, vol. 21, pp. 768–769, 1965.

[34] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comp. Applied Math.*, vol. 20, pp. 53–65, 1987.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[36] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[37] Y. Ban, G. Rosman, T. Ward, D. Hashimoto, T. Kondo, O. Meireles, and D. Rus, "Aggregating long-term context for learning surgical workflows," *arXiv preprint arXiv:2009.00681*, 2020.

[38] J. Zhang, Y. Nie, Y. Lyu, H. Li, J. Chang, X. Yang, and J. J. Zhang, "Symmetric dilated convolution for surgical gesture recognition," *arXiv preprint arXiv:2007.06373*, 2020.